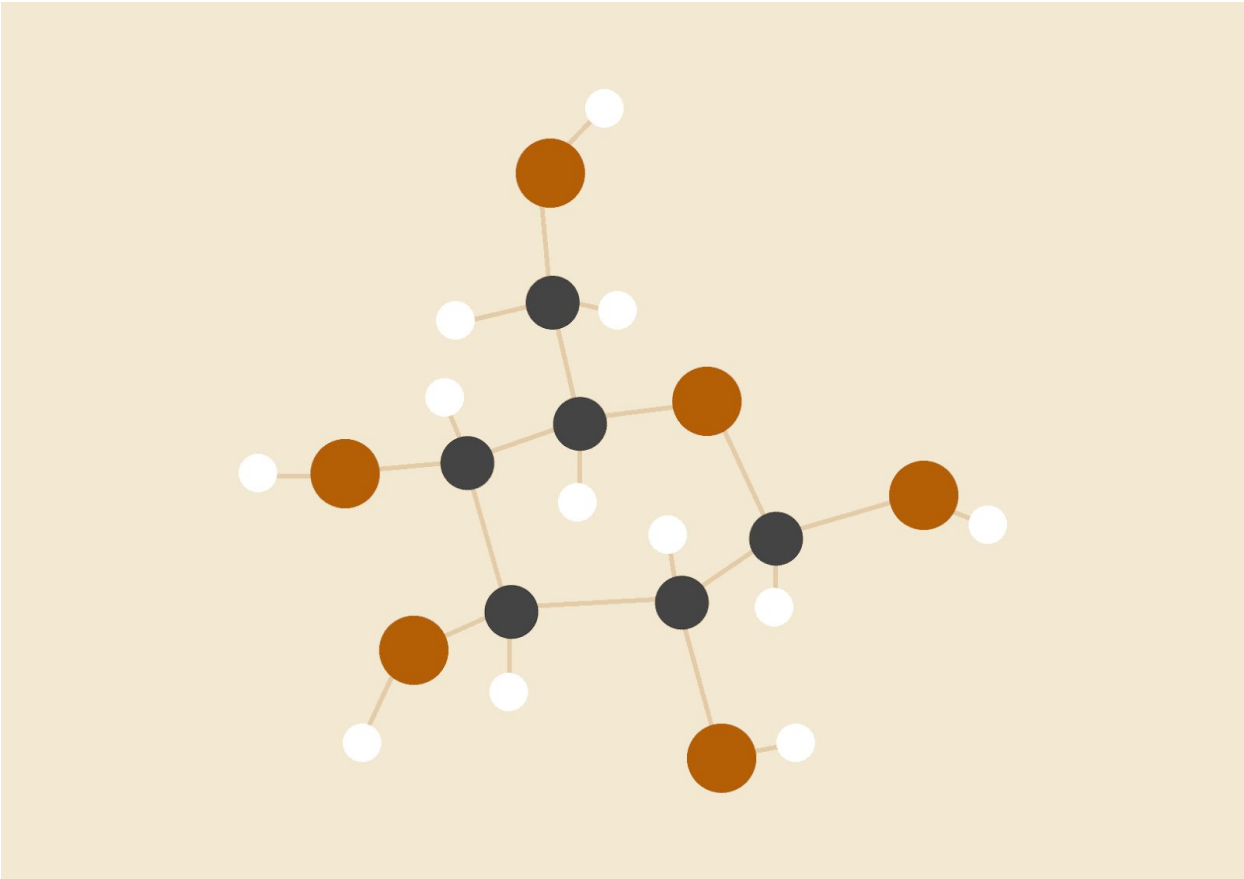
Wine Quality Analysis



A Sai Akhilesh

Indian Institute of Information Technology, Sricity.

PROBLEM STATEMENT

White vinho verde wine dataset is given. Model the wine quality based on input variables obtained from physicochemical tests.

The problem demands for regression modelling.

ABSTRACT

This report gives insights about the white-wine data in the form of metrics and visualizations. Performed Regression Modelling, Adequacy checks, Model diagnostics.

The goodness of fit is measured by R-squared metric.

CONTENTS

INTRODUCTION

The first step in every data science project is to get some minimal understanding of the domain. I have gone through some websites to understand different attributes that describe the quality of the Wine.

Understanding Wine and Types:

1. Wine is an alcoholic beverage made from grapes which is fermented without the addition of

sugars, acids, enzymes, water, or other nutrients.

2. White wine is made from white grapes with no skins or seeds

Understanding Wine Attributes and Properties:

1. Acidity: Measured in grams/dm3.

Acid types are :

1. Fixed acidity
2. Volatile Acidity
3. Citric acidity
4. pH is used to define the acidity

2. Sweetness: Measured in grams/dm3.

Sweetness measure is given by Residual Sugar.

3.Salt Measure: Measured in grams/dm3.

Salty measure is given by Chlorides and Sulphites.

Sulphite types are:

1. Sulphates
2. Free sulfur dioxide
3. Total sulfur dioxide

4. Alcohol Measure: Measured in % vol. Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine.

EXPLORATORY DATA ANALYSIS (EDA)

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness: A measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

If skewness is less than -1 or greater than 1, the distribution is highly skewed. If between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.

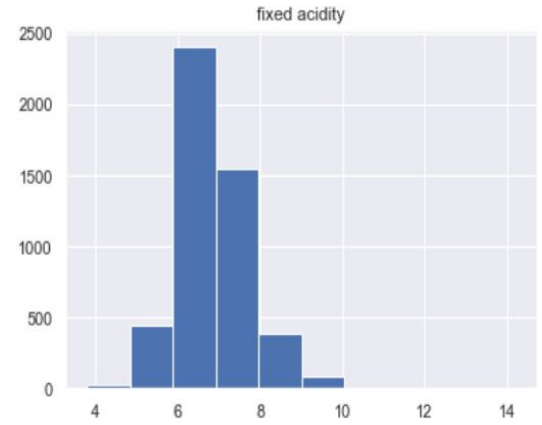
Kurtosis: A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

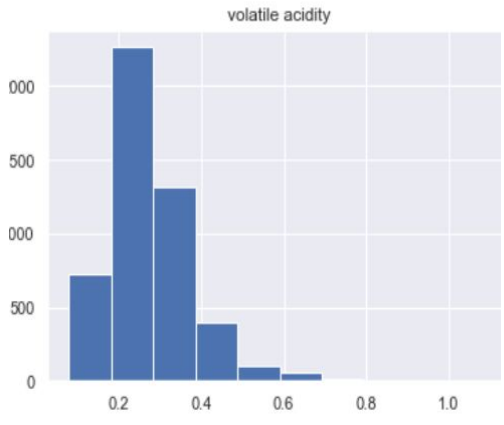
For any univariate normal distribution kurtosis is 3.

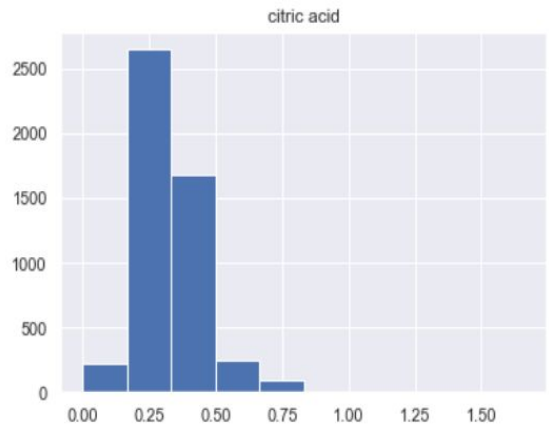
The table also shows skewness and kurtosis values of each variable.

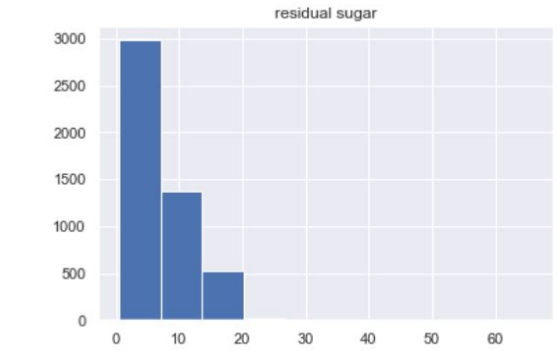
	dtype	distincts	nulls	missing ration	uniques	skewness	kurtosis
fixed acidity	float64	68	0	0.0	[[7.0, 6.3, 8.1, 7.2, 6.2, 8.6, 7.9, 6.6, 8.3,...	0.647751	2.172178
volatile acidity	float64	125	0	0.0	[[0.27, 0.3, 0.28, 0.23, 0.32, 0.22, 0.18, 0.1...	1.576980	5.091626
citric acid	float64	87	0	0.0	[[0.36, 0.34, 0.4, 0.32, 0.16, 0.43, 0.41, 0.3...	1.281920	6.174901
residual sugar	float64	310	0	0.0	[[20.7, 1.6, 6.9, 8.5, 7.0, 1.5, 1.45, 4.2, 1....	1.077094	3.469820
chlorides	float64	160	0	0.0	[[0.045, 0.049, 0.05, 0.057999999999999996, 0....	5.023331	37.564600
free sulfur dioxide	float64	132	0	0.0	[[45.0, 14.0, 30.0, 47.0, 28.0, 11.0, 17.0, 16...	1.406745	11.466342
total sulfur dioxide	float64	251	0	0.0	[[170.0, 132.0, 97.0, 186.0, 136.0, 129.0, 63....	0.390710	0.571853
density	float64	890	0	0.0	[[1.001, 0.9940000000000001, 0.9951, 0.9956, 0...	0.977773	9.793807
pH	float64	103	0	0.0	[[3.0, 3.3, 3.26, 3.19, 3.18, 3.22, 2.99, 3.14...	0.457783	0.530775
sulphates	float64	79	0	0.0	[[0.45, 0.49, 0.44, 0.4, 0.47, 0.56, 0.53, 0.6...	0.977194	1.590930
alcohol	float64	103	0	0.0	[[8.8, 9.5, 10.1, 9.9, 9.6, 11.0, 12.0, 9.7, 1...	0.487342	-0.698425
quality	int64	7	0	0.0	[[6.0, 5.0, 7.0, 8.0, 4.0, 3.0, 9.0]]	0.155796	0.216526

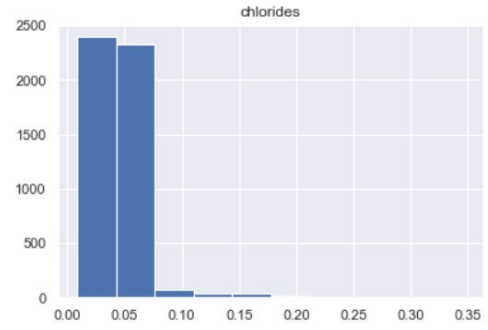
The distribution of each independent variable (feature) in order is given below.

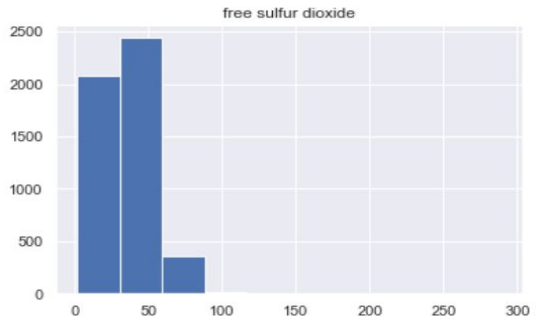


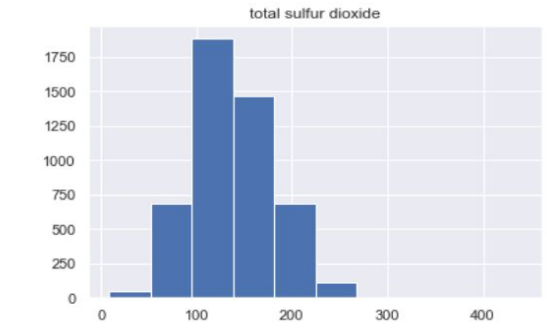


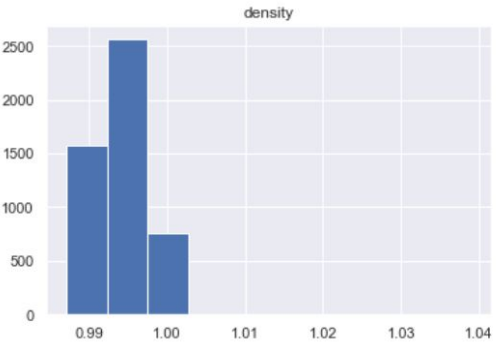


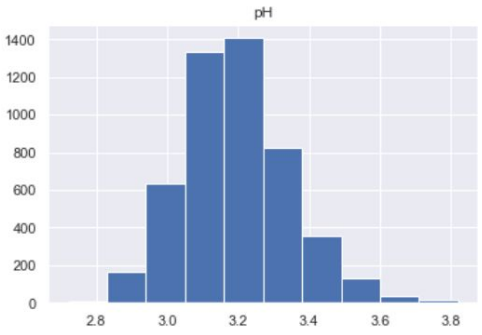


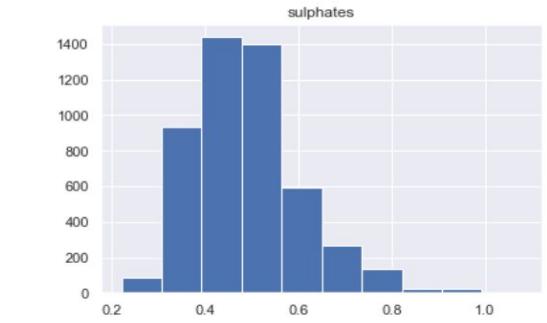


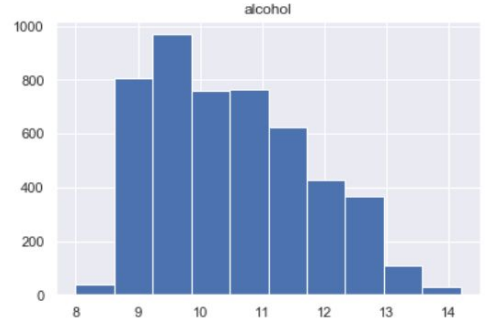




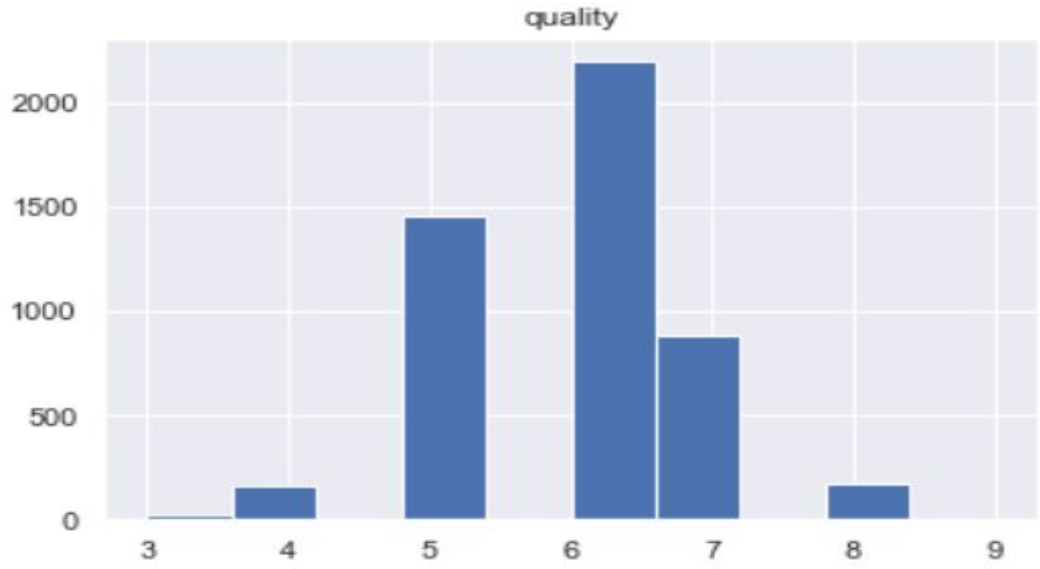








The distribution of the dependent variable (target) - *'Quality'*



I. Data Pre-Processing

1. **Missing Values :** There are no missing values in the given Dataset.

```
Shape: (4898, 12)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
None
```

2. **Shuffling :** I have shuffled the Dataset.

3. **Balancing :** We can see that most of the wines being of quality 6. While very few wines have quality 3, 9. This tells that the dataset is an imbalanced one.
4. **Scaling:** I have standardized the Data so that all variables are on the same scale and equally important.

II. Multiple Linear Regression

I have applied Ordinary Least Squares method to estimate the parameters of the regression line.

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.282
Model:	OLS	**Adj. R-squared:**	0.280
Method:	Least Squares	**F-statistic:**	174.3
Date:	Mon, 08 Jul 2019	**Prob (F-statistic):**	0.00
Time:	13:07:14	**Log-Likelihood:**	-5543.7
No. Observations:	4898	**AIC:**	1.111e+04
Df Residuals:	4886	**BIC:**	1.119e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8779	0.011	547.502	0.000	5.857	5.899
fixed acidity	0.0553	0.018	3.139	0.002	0.021	0.090
volatile acidity	-0.1878	0.011	-16.373	0.000	-0.210	-0.165
citric acid	0.0027	0.012	0.231	0.818	-0.020	0.025
residual sugar	0.4132	0.038	10.825	0.000	0.338	0.488
chlorides	-0.0054	0.012	-0.452	0.651	-0.029	0.018
free sulfur dioxide	0.0635	0.014	4.422	0.000	0.035	0.092
total sulfur dioxide	-0.0121	0.016	-0.756	0.450	-0.044	0.019
density	-0.4494	0.057	-7.879	0.000	-0.561	-0.338
pH	0.1036	0.016	6.513	0.000	0.072	0.135
sulphates	0.0721	0.011	6.291	0.000	0.050	0.095
alcohol	0.2381	0.030	7.988	0.000	0.180	0.297

Omnibus:	114.161	**Durbin-Watson:**	2.027
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	251.637
Skew:	0.073	**Prob(JB):**	2.28e-55
Kurtosis:	4.101	**Cond. No.**	12.5

We observe that We

observe that R-squared is 0.282.

It means the Linear Regression model is able to explain only 28.2% of the total variability of the Data. It is quite low.

Possible Reasons:

1. The Data consists of only 4898 samples and might be having many outliers. Hence, it might have not captured the total variability.

Let's explore the Distributions of each variable for more insights (shown above).

We observe that the distributions of all variables are not properly distributed. For example, the variables like Residual Sugar, Chlorides, Free Sulfur Dioxide and Density are having many samples away from the general mass.

Let's remove 10% of samples of each variable on either side of the Distribution and see if we can fit a better model.

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.206
Model:	OLS	**Adj. R-squared:**	0.203
Method:	Least Squares	**F-statistic:**	67.46
Date:	Mon, 08 Jul 2019	**Prob (F-statistic):**	1.37e-134
Time:	13:07:25	**Log-Likelihood:**	-3267.4
No. Observations:	2871	**AIC:**	6559.
Df Residuals:	2859	**BIC:**	6630.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.0258	0.014	426.671	0.000	5.998	6.053
fixed acidity	0.0982	0.022	4.549	0.000	0.056	0.141
volatile acidity	-0.1256	0.015	-8.247	0.000	-0.155	-0.096
citric acid	-0.0714	0.015	-4.919	0.000	-0.100	-0.043
residual sugar	0.4393	0.059	7.433	0.000	0.323	0.555
chlorides	-0.0246	0.017	-1.470	0.142	-0.057	0.008
free sulfur dioxide	0.0749	0.019	4.027	0.000	0.038	0.111
total sulfur dioxide	-0.0195	0.021	-0.913	0.362	-0.061	0.022
density	-0.5403	0.093	-5.793	0.000	-0.723	-0.357
pH	0.1109	0.022	5.129	0.000	0.068	0.153
sulphates	0.0639	0.015	4.159	0.000	0.034	0.094
alcohol	0.1052	0.049	2.138	0.033	0.009	0.202

Omnibus:	47.943	**Durbin-Watson:**	1.584
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74.809
Skew:	0.159	**Prob(JB):**	5.69e-17
Kurtosis:	3.724	**Cond. No.**	15.7

We observe that the R-squared value has come down to 20.6%

It means that, now the model is able to explain only 20.6% of the total variability.

Possible Reasons:

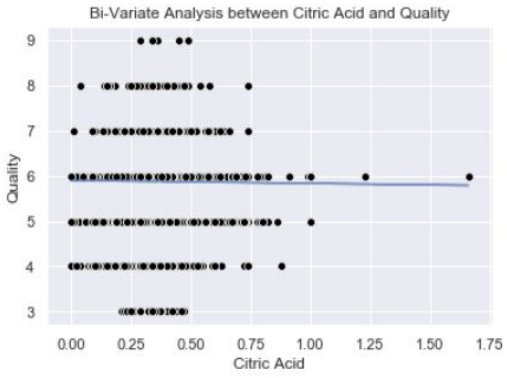
1. When we removed the extreme 10% samples for each variable, 2250 samples were removed as a whole. It means we lost almost 50% of the Data.
2. It might also be possible that the variable 'Quality' depends on the variables which are correlated.

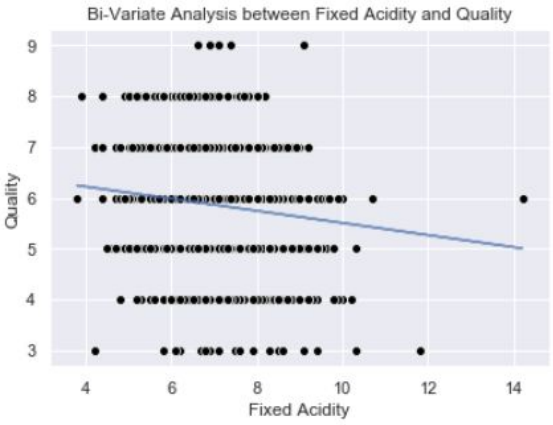
Let's analyse the assumptions of Multiple Linear Regression to get more insights.

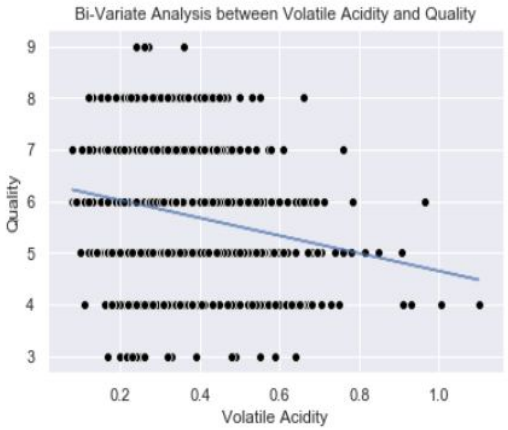
MULTIPLE LINEAR REGRESSION (MLR) - TEST OF ASSUMPTIONS

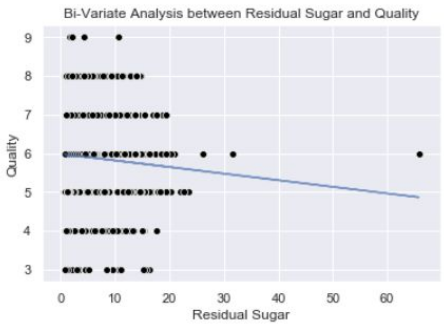
1. **Linearity of Independent variables:** Each independent variable must have a linear relationship with the Dependent Variable.

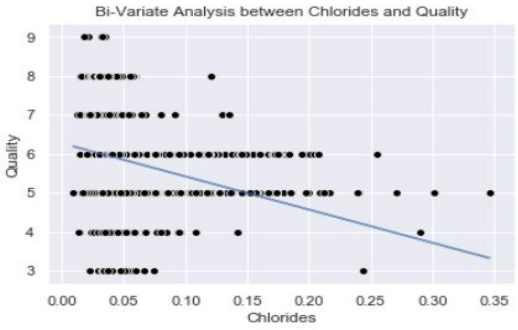
The relationship of each independent variable with the Dependent variable (Quality) in order is given below

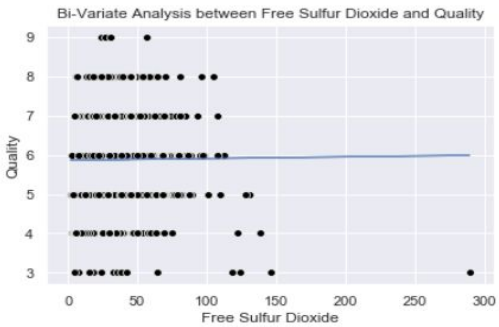


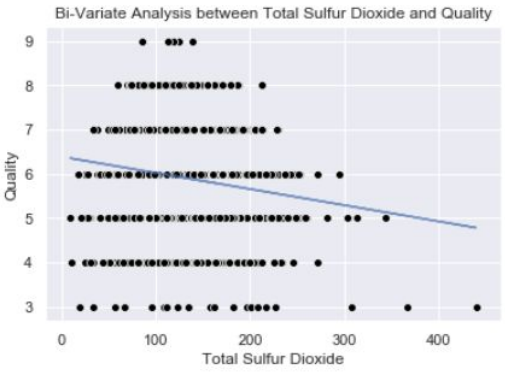


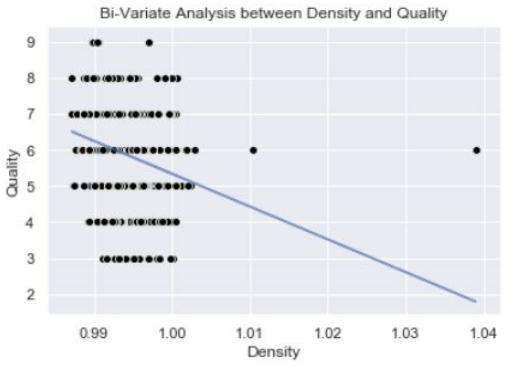


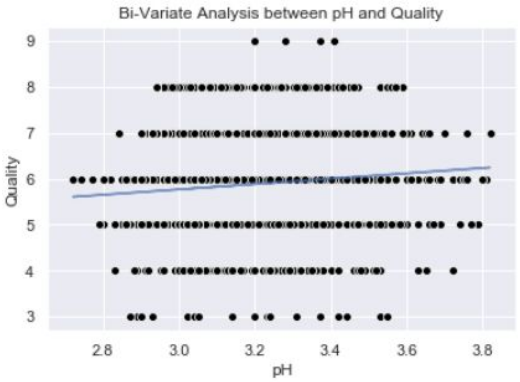


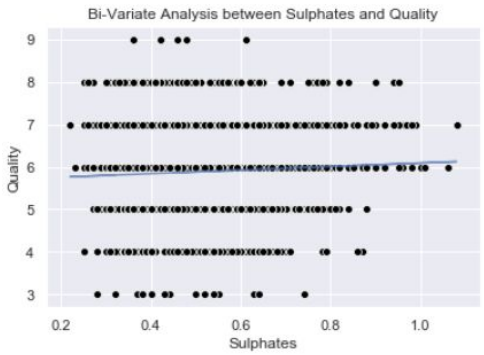


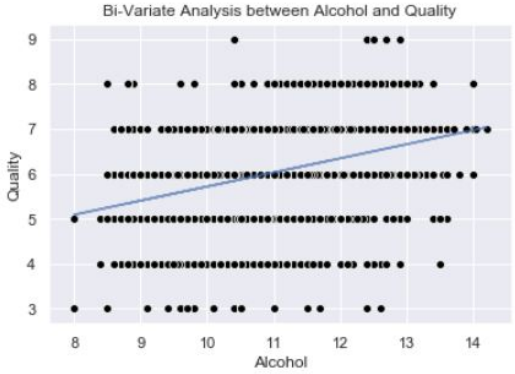




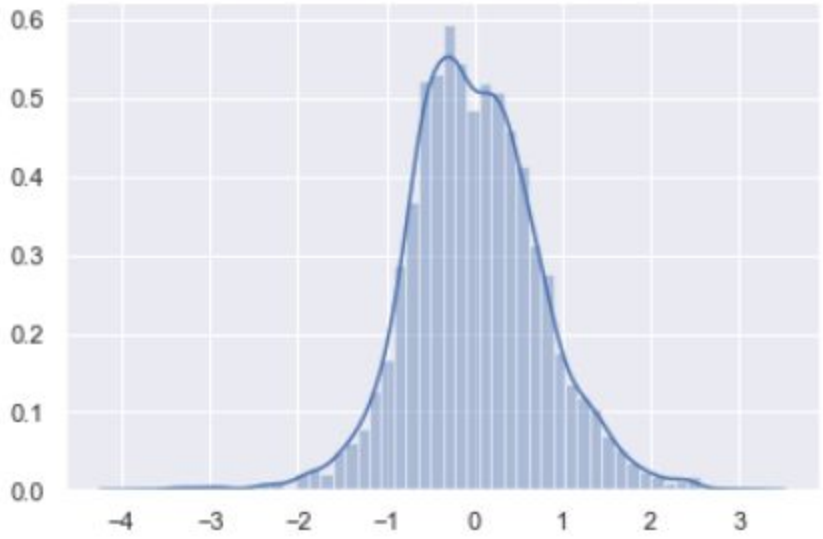




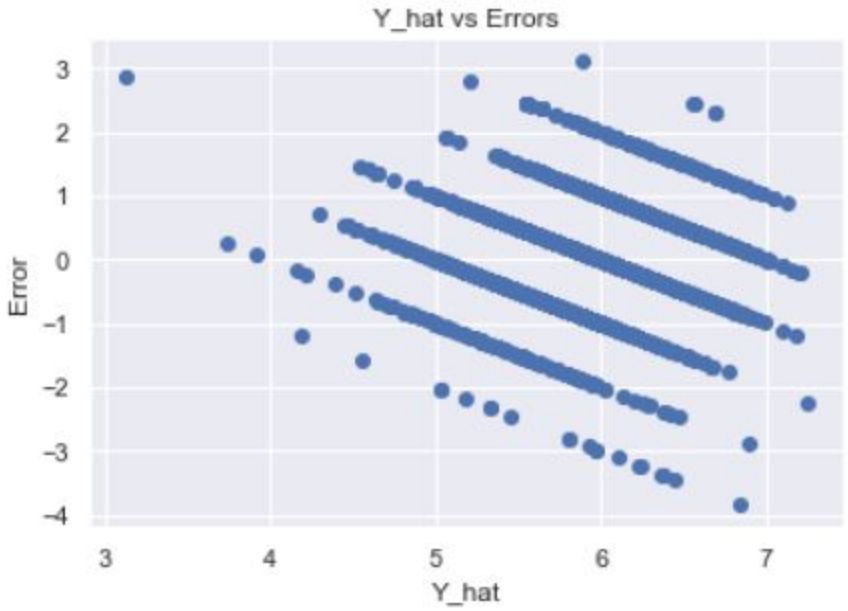




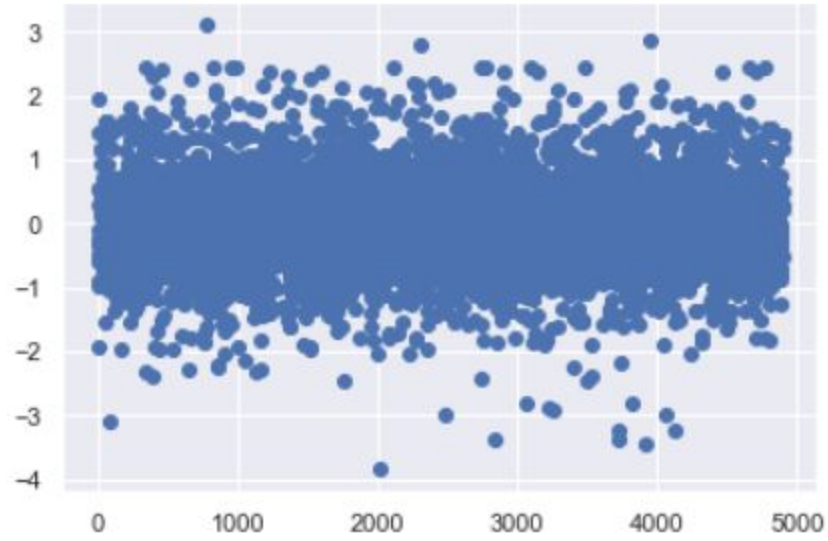
2. **<u>Normal Distribution of Residuals(Errors):</u>** The errors i.e, the differences between the predicted and target values must be normally distributed.



3. **Homoskedasticity Condition:** Equal variance of Residuals.



4. **Uncorrelated Residuals:**



Observations:

1. There is no strong linear relationship of the independent variables with the dependent variable. This might be a major possible reason why our linear model has captured low variability.

Remedies:

1. The data must be transformed so that each independent variable is linearly dependent with the Dependent Variable.
2. We have to apply Non-Linear Methods which can capture non-linear Data.

MODEL DIAGNOSTICS:

1. **Test of individual Parameters:**

   Hypothesis Testing:

   Let alpha(level of significance) = 0.05, If p-value <= 0.05, we can reject the null hypothesis.

   H 0 : β = 0 (where β is coefficient of the variable in the fitted line)

   H1: β != 0

   When applied Linear Regression separately for each variable, we obtained p-values > 0.05 only for the variables 'citric acid' and 'free sulfur dioxide'.

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.4153
Date:	Mon, 08 Jul 2019	Prob (F-statistic):	0.519
Time:	20:33:44	Log-Likelihood:	-6354.4
No. Observations:	4898	AIC:	1.271e+04
Df Residuals:	4896	BIC:	1.273e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8779	0.013	464.462	0.000	5.853	5.903
citric acid	-0.0082	0.013	-0.644	0.519	-0.033	0.017

Omnibus:	27.428	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.026
Skew:	0.154	Prob(JB):	4.98e-07
Kurtosis:	3.217	Cond. No.	1.00

Warnings:

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.3259
Date:	Mon, 08 Jul 2019	Prob (F-statistic):	0.568
Time:	20:33:44	Log-Likelihood:	-6354.5
No. Observations:	4898	AIC:	1.271e+04
Df Residuals:	4896	BIC:	1.273e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8779	0.013	464.457	0.000	5.853	5.903
free sulfur dioxide	0.0072	0.013	0.571	0.568	-0.018	0.032

Omnibus:	27.869	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.413
Skew:	0.157	Prob(JB):	4.10e-07
Kurtosis:	3.214	Cond. No.	1.00

It means they have no significant predictive power.(Their p-value are much higher than 0.05). We can safely remove such variables and refit the model

After removing these two variables, we have 9(11-2) independent variables.

Fitting MLR again we get,

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.279
Model:	OLS	**Adj. R-squared:**	0.278
Method:	Least Squares	**F-statistic:**	210.1
Date:	Mon, 08 Jul 2019	**Prob (F-statistic):**	0.00
Time:	13:07:31	**Log-Likelihood:**	-5553.6
No. Observations:	4898	**AIC:**	1.113e+04
Df Residuals:	4888	**BIC:**	1.119e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8779	0.011	546.513	0.000	5.857	5.899
fixed acidity	0.0561	0.018	3.202	0.001	0.022	0.090
volatile acidity	-0.1982	0.011	-17.872	0.000	-0.220	-0.176
residual sugar	0.4428	0.038	11.776	0.000	0.369	0.517
chlorides	-0.0033	0.012	-0.282	0.778	-0.027	0.020
total sulfur dioxide	0.0304	0.013	2.353	0.019	0.005	0.056
density	-0.4873	0.056	-8.649	0.000	-0.598	-0.377
pH	0.1068	0.016	6.731	0.000	0.076	0.138
sulphates	0.0728	0.011	6.347	0.000	0.050	0.095
alcohol	0.2260	0.030	7.640	0.000	0.168	0.284

Omnibus:	104.354	**Durbin-Watson:**	2.024
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	217.875
Skew:	0.081	**Prob(JB):**	4.89e-48
Kurtosis:	4.021	**Cond. No.**	11.8

2. **Multicollinearity:**

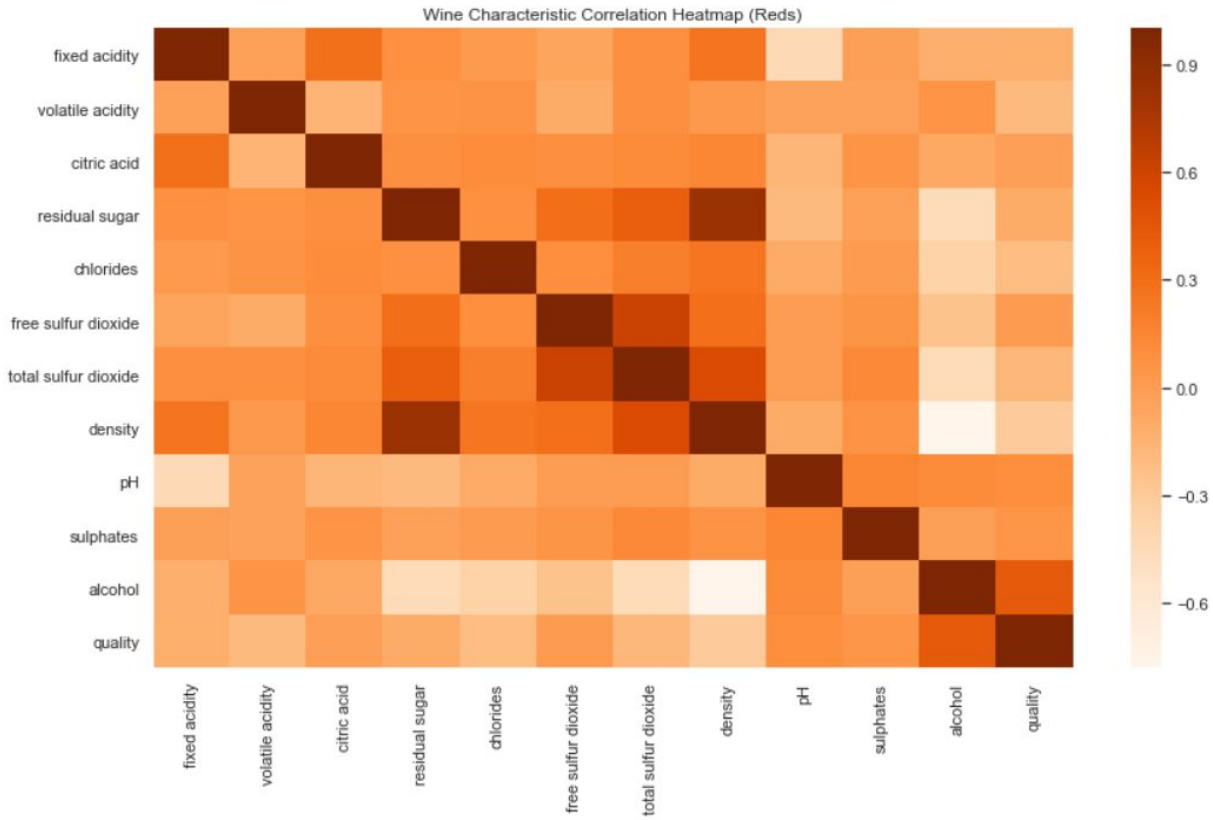
Correlation Matrix: To know the correlation between the variables, I plotted heatmap. Finding correlation helps in getting the following insights:

● If there is a strong correlation between some of the independent variables, we can do preprocessing like

selecting important variable among them/ linear combination of such variables, etc. This helps in overcoming the multicollinearity problem.

● If there is not much correlation between some of the independent variables with the dependent variable, we

must perform statistical tests to decide whether to include such variables in the model or not.



Observations:

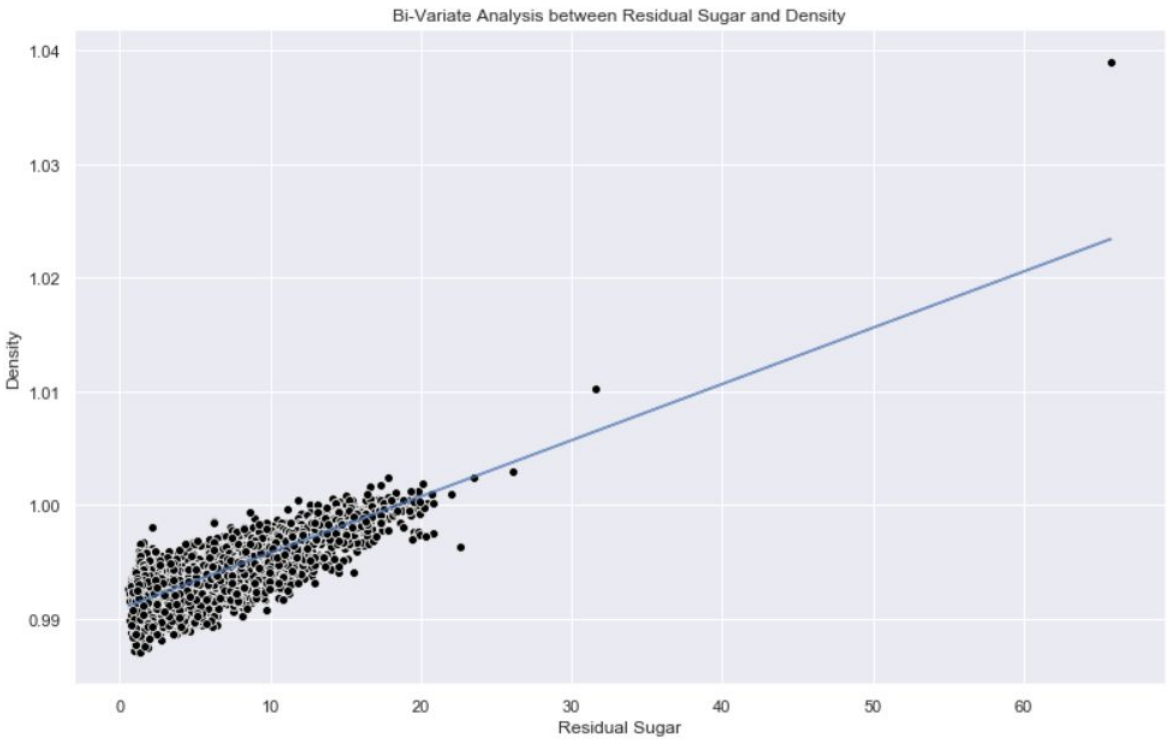
1. There is a strong positive correlation between density and residual sugar.

2. There is a strong negative correlation between density and alcohol.

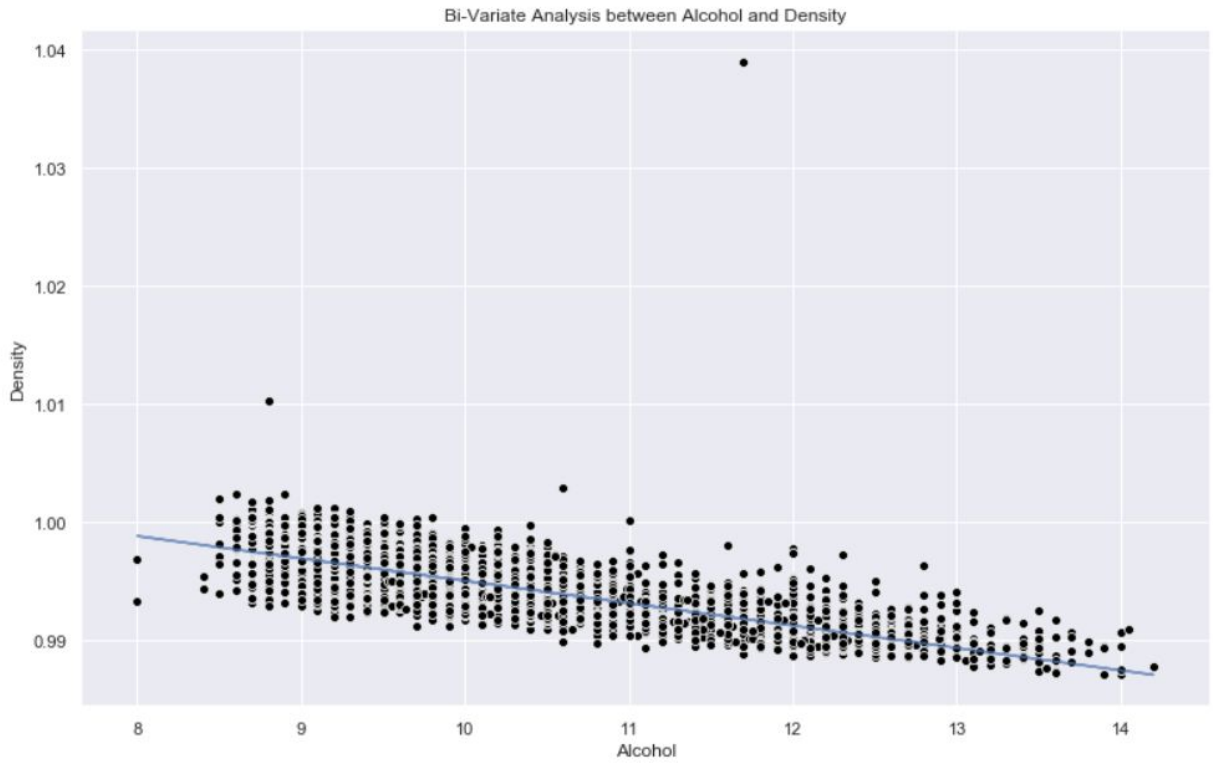
3. There is no significant correlation between other independent variables.

4. Correlation between quality is high with alcohol than other independent variables

Bivariate Analysis between Density and Residual Sugar



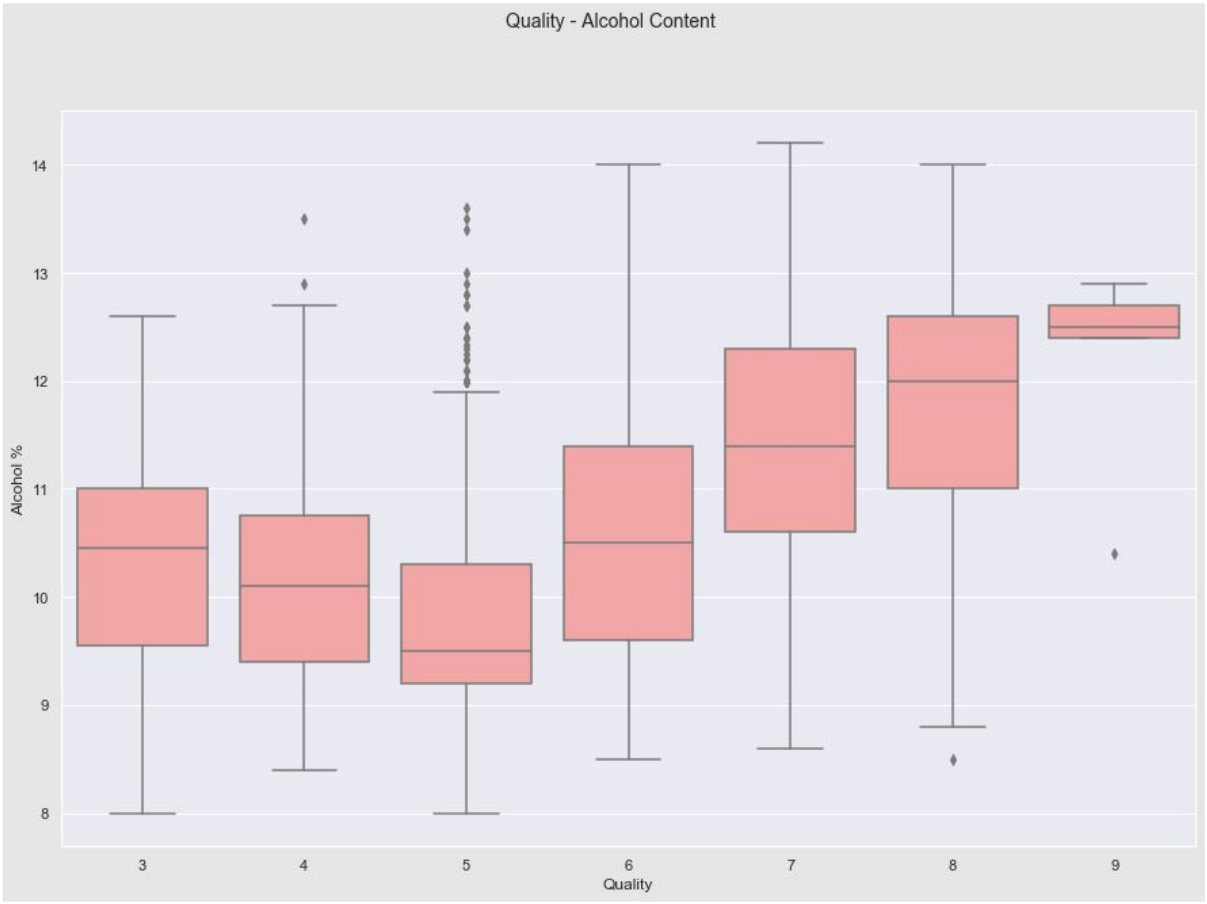
Bivariate Analysis between Density and Alcohol



Observations:

1.Bivariate scatter plot shows that with increase in Residual Sugar content, Density also increases, supporting the results given by the correlation matrix (+ve correlation).

2.The second graph shows that with increase in Alcohol content, Density decreases, supporting the results given by the correlation matrix(-ve correlation).



Observations:

1. From the above graph, we can see that there is a slight increasing trend between Alcohol% and Quality.

2. Generally, wines with high alcohol concentration are rated high in quality. In contrast, we can also observe that there are some outliers in the third box-plot (for quality 5).

3. Even when alcohol percentage is high, some wines are labeled as quality 5(not high). Since alcohol% is not the only factor, may be because of the other independent variables, overall quality of the wine is reduced.

VIF (Variance Inflation Factor):

Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression. It is may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the R-squared from that regression. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone [1/(1-R2)].

	VIF	Features
0	1.000000	const
2	1.063069	volatile acidity
8	1.136475	sulphates
4	1.221320	chlorides
5	1.442547	total sulfur dioxide
7	2.176103	pH
1	2.650908	fixed acidity
9	7.567793	alcohol
3	12.222682	residual sugar
6	27.441035	density

Observations:

1.We saw both in the form of visualization (Bivariate Graphs) and Correlation matrix that Density has strong positive correlation with Residual Sugar and strong negative correlation with Alcohol %.`

2.Hence including such variable in linear regression model would violate the multicollinearity assumption.

Remedies:

1.Remove highly correlated variables

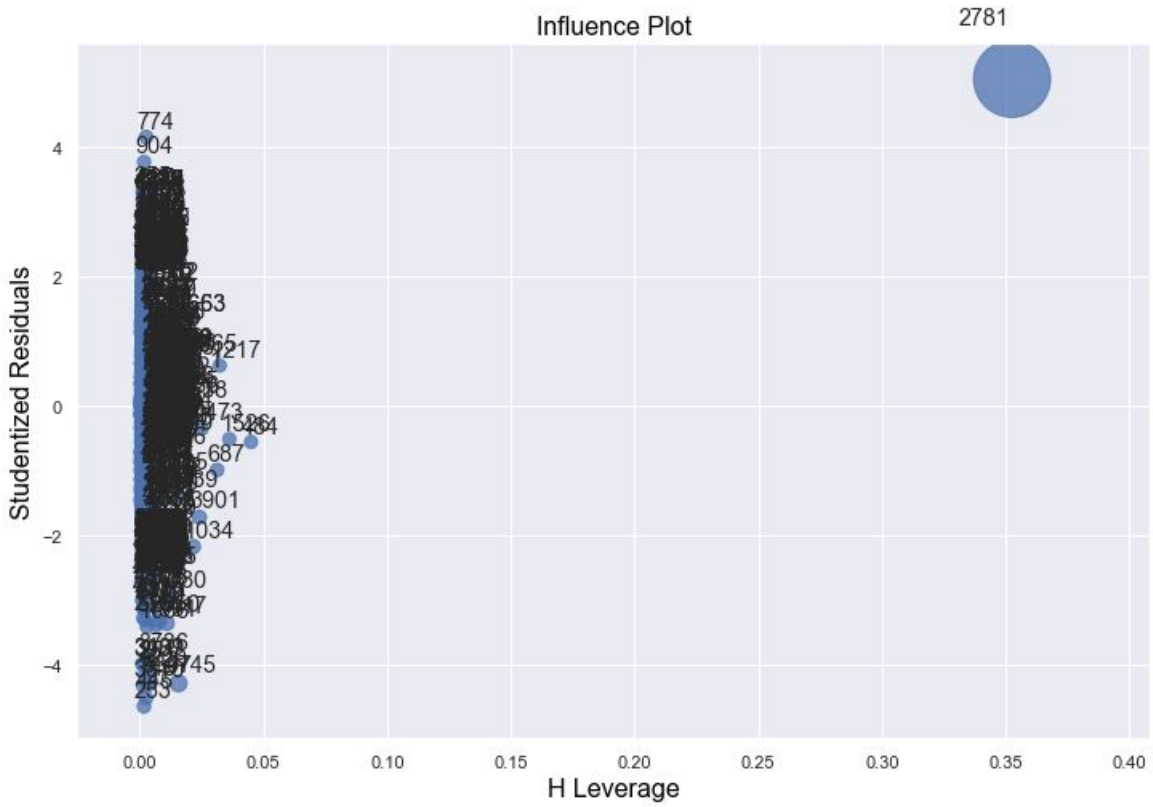
2.Use linear combination of correlated variables/ apply PCA.

3. **Outliers, Leverage Points and Influential Points:**

Outliers: Observations which lie much away from the general mass related to y.

Leverage Points: Observations which lie much away from the general mass related to x.

Influence Points: If any of outliers or leverage points influence the regression estimates, they are known as influential points.



Sample with index number 2781 is the only influential point. I have removed it.

PCA - PRINCIPAL COMPONENT ANALYSIS

Since the data has only 9 independent variables, removing some of them will result in loss of information.

Hence, I have applied PCA to overcome multicollinearity problem and it also helps in retaining much of the information. Following are the results:

Principal Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Cumulative Variance Explained	0.329	0.486	0.604	0.712	0.820	0.889	0.950	0.997	1.

Now, after transforming Multicollinearity Problem is solved.

	VIF	Features
0	1.0	const
1	1.0	PC1
2	1.0	PC2
3	1.0	PC3
4	1.0	PC4
5	1.0	PC5
6	1.0	PC6
7	1.0	PC7
8	1.0	PC8
9	1.0	PC9

Data Transformation:

After transforming the data using PCA, I applied power transformation and fitted the OLS model.

Power transformation is done to make distribution of the data closer to normal distribution.Yeo-johnson is used for the transformation. R-squared value slightly increased from 0.279 to 0.282.

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.282
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	213.4
Date:	Tue, 09 Jul 2019	**Prob (F-statistic):**	0.00
Time:	01:39:00	**Log-Likelihood:**	-5543.1
No. Observations:	4898	**AIC:**	1.111e+04
Df Residuals:	4888	**BIC:**	1.117e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8146	0.011	524.327	0.000	5.793	5.836
PC1	-0.3098	0.011	-28.555	0.000	-0.331	-0.289
PC2	-0.0213	0.011	-1.980	0.048	-0.042	-0.000
PC3	-0.1677	0.011	-15.034	0.000	-0.190	-0.146
PC4	-0.1014	0.011	-8.849	0.000	-0.124	-0.079
PC5	-0.0692	0.011	-6.288	0.000	-0.091	-0.048
PC6	0.0452	0.011	4.197	0.000	0.024	0.066
PC7	-0.1117	0.011	-10.279	0.000	-0.133	-0.090
PC8	0.2528	0.011	23.283	0.000	0.231	0.274
PC9	-0.0949	0.012	-7.635	0.000	-0.119	-0.071

Omnibus:	101.865	**Durbin-Watson:**	1.623
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	217.540
Skew:	0.047	**Prob(JB):**	5.78e-48
Kurtosis:	4.028	**Cond. No.**	1.35

CONCLUSIONS:

1. For the given data, OLS method was used to estimate the parameters. Citric acid, free sulphur dioxide variables gave p-value more than 0.05 and hence were not used for quality prediction.

2. Multicollinearity problem was solved by applying PCA on the data.

3. Heteroskedasticity plot shows that the variance is constant and QQ-plot shows that the error terms are normally distributed. The best value of R-square using linear model is 0.282.

4. Non-linear models can be applied on the data to check if the prediction power is better than linear regression.

5. There might some other variables like Grape Ripening, Temperature, Brand etc which might be impacting this quality variable.